Fuzzy Cluster Analysis of Bioinformatics Data Composed of Microarray Expression Data and Gene Ontology Annotations

Timothy C. Havens, James M. Keller, Mihail Popescu, James C. Bezdek
Dept. of Electrical and Computer Engineering
University of Missouri
Columbia, MO 65211, USA
havenst@gmail.com, kellerj@missouri.edu, popescum@missouri.edu, jbezdek@uwf.edu

Erin MacNeal Rehrig, Heidi M. Appel, Jack C. Schultz
Bond Life Sciences Center
University of Missouri
Columbia, MO 65211, USA
rehrige@missouri.edu, appelh@missouri.edu, schultzjc@missouri.edu

***Abstract*—This paper presents the framework and results of the cluster analysis of a selected set of Arabidopsis (a leafy plant) genes in the presence of insect-feeding and wounding stress. We outline the methodology by which we coupled the results of a microarray experiment with the Gene Ontology (GO) annotations of each gene to produce aggregate relational data. Our method combines two relational matrices: one matrix is derived from a fuzzy GO similarity measure and another is derived from the microarray data using a statistical similarity measure. Finally, we used a fuzzy clustering algorithm (NERFcM) and a validity measure (CCV) to cluster and validate the resulting relational data. Results are presented that outline the functional summarization of the clusters. The methods presented here give microarray researchers additional tools to investigate relations between gene expression and gene functions.**

I. Introduction

Bioinformatics data can take many forms: e.g. patient records, ontology annotations, and microarray experiments. Microarray experiments allow biologists to characterize the expression of a group of genes in the presence of treatments, such as drugs, toxic chemicals, or stresses. These experiments provide information on how genes express relative to each other. Hence, *functional annotation* can be performed by inference from well-characterized genes [1, 2]. On the other hand, if a group of similarly behaving genes are already well annotated, these annotations can be used to produce a *functional summarization* of the group. In this paper, we couple methods used in both functional annotation and functional summarization to perform a novel analysis of a set of plant genes. We achieve this by combining the results of a microarray experiment with *Gene Ontology* (GO) annotations. Figure 1 is an illustration of our procedure.

The GO is a hierarchical taxonomy of functional annotations of genes [3]. Each term in the GO is taken from a controlled vocabulary, or corpus, and describes gene and gene product attributes. Suppose two genes, G_1 and G_2, are represented by a set of GO terms $G_1 = \{T_{11}, T_{12}, \ldots, T_{1n}\}$ and $G_2 = \{T_{21}, T_{22}, \ldots, T_{2m}\}$. For these sets, we can compute a similarity value using a number of methods [see 4–8]. We use an *augmented fuzzy measure-based similarity* (AFMS) as described in references [6, 7]. This similarity measure has been shown to overcome limitations that are present in both pair-wise aggregation methods, set-based similarity measures, and vector space-based similarity measures (the best known being the vector cosine) [6, 7]. Given a set of gene products $\{G_1, G_2, \ldots, G_N\}$, we calculate an $N \times N$ dissimilarity matrix $\mathbf{D}$, where D_{ij} is the dissimilarity value of G_i and G_j.

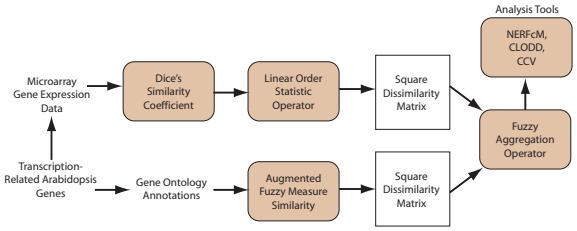


Fig. 1. Fuzzy cluster analysis block diagram

DNA microarrays (also called gene or genome chips, DNA chips, and gene arrays) measure gene expression level in the presence of a treatment. Detailed information about gene expression and microarrays can be found in references [9, 10]. In brief, microarrays measure expression level by monitoring fluorescent emission at array spots. The fluorescent emission is proportional to the expression of the corresponding gene (fragment). We examined a microarray experiment that measured the expression levels of 3,044 *Arabidopsis thaliana* genes that are known to be related to *transcription factors*—proteins that bind to regulatory regions and helps control gene expression. The treatments applied to the plants were insect-feeding and wounding stresses. In this paper, we focus on 198 of the 3,044 genes, denoted as TF_{198}. These genes were selected for their important role in regulating gene expression and responsiveness to stress. Section II describes this data in detail.

We began by calculating relational data for each data set

978-1-4244-2352-1/08/$25.00 ©2008 IEEE

TABLE I
INSECT-FEEDING MICROARRAY EXPERIMENT DATA EXAMPLE.

Insect	*Pieris*				*Spodoptera*				*Brevicoryne*		*Myzus*		Wounding			
Tissue	Local		Systemic		Local		Systemic		Local		Local		Local		Systemic	
Sample Time (hrs)	6	24	6	24	6	24	6	24	6	24	6	24	6	24	6	24
GO id																
AT2G35700					UP	UP										
AT1G46768											DOWN					
AT4G17710									UP							
AT5G60890	UP				UP				DOWN		DOWN					

(GO annotations and microarray expression). Hence, each pair of genes will have two (dis)similarity values, one based on the GO annotations, and one based on the microarray experiment. Section III describes these relational data. We then combine these relational data using a fuzzy aggregation operator. Finally, a relational clustering algorithm and validation method (NERFcM and CCV [11, 12]) were used to partition these aggregate data. Section IV describes the methodology and results of this analysis. We wrap up this paper in Section V.

II. ARABIDOPSIS DATA

The *Arabidopsis thaliana* plant is unique in that it was the first plant genome to be sequenced [13]. As a result, it is widely used in plant sciences, especially in genetics. Each plant was exposed to one of four different insects, the *Pieris rapae* (cabbageworm), the *Spodoptera exigua* (beet armyworm), the *Brevicoryne brassicae* (cabbage aphid), and the *Myzus persicae* (green peach aphid). As a stress control, a set of plants were also exposed to mechanical wounding as a stress control. The caterpillars were allowed to feed until 10-30% of the leaf area was eaten (about 2-4 hours). The caterpillars were then removed. Aphids have effects on plants that are weaker and slower to develop than those of caterpillars, thus, aphids were allowed to feed for one week and then removed. Gene expression was measured at six and 24 hours following insect removal in local tissue (treated leaf) and systemic tissue (untreated leaf). Systemic tissue measurements were not possible on the aphid-treated plants as the aphids were too small to be localized. Gene expression was measured with a whole genome Operon oligo microarray (v1) with 64 hybridizations of treatment and control RNA for four biological replicates for each treatment and time. The resulting data were filtered with a two-fold expression ratio cutoff and analyzed by ANOVA. 3,044 genes were differentially expressed in response to the treatment, of which 198 were transcription factors. Table I shows examples of the expression data for four of the TF_{198} genes. **UP** indicates the gene *UP-regulated* in the presence of the treatment, **DOWN** indicates *DOWN-regulation*, and no entry indicates no expression. As the table shows, the data is sparse and there is little mixing of expression values between insects. This is representative of the entire data set.

Additionally, the GO annotations of TF_{198} genes were downloaded from the Arabidopsis Information Resource (TAIR), which is a database of genetic data for the Arabidopsis plant [13]. Table II shows the GO annotations for the four genes shown in Table I. This table illustrates the main problem with using only GO annotations to compute relational data. While human gene products are well annotated, the Arabidopsis has many identical entries and the annotations tend to be very general. For example, the GO terms GO:0003700-*transcription factor activity* and GO:0005634-*nucleus* provide very little information content as they are used to annotate virtually *all* the genes in TF_{198}. Furthermore, Table II shows that AT2G35700, AT1G46768, and AT4G17710 have identical GO annotations; however, they have distinctly different expression data. Hence, combining the microarray expression data with the GO annotation data should provide a synergistic view into the inter-relationships of the genes and their functions.

TABLE II
EXAMPLE GO ANNOTATIONS FOR ARABIDOPSIS TRANSCRIPTION FACTOR-RELATED GENES.

GO id	GO term	Definition
AT2G35700	GO:0003677	DNA binding
	GO:0003700	Transcription factor activity
	GO:0005634	Nucleus
	GO:0006355	Regulation of transcription, DNA-dep.
AT1G46768	GO:0003677	DNA binding
	GO:0003700	Transcription factor activity
	GO:0005634	Nucleus
	GO:0006355	Regulation of transcription, DNA-dep.
AT4G17710	GO:0003677	DNA binding
	GO:0003700	Transcription factor activity
	GO:0005634	Nucleus
	GO:0006355	Regulation of transcription, DNA-dep.
AT5G60890	GO:0000162	Tryptophan biosynthesis
	GO:0003677	DNA binding
	GO:0003700	Transcription factor activity
	GO:0005634	Nucleus
	GO:0009651	Response to salt stress
	GO:0009737	Response to abscisic acid stimulus
	GO:0009739	Response to gibberellic acid stimulus
	GO:0009751	Response to salicylic acid stimulus
	GO:0009753	Response to jasmonic acid stimulus
	GO:0009759	Indole glucosinolate biosynthesis
	GO:0016301	Kinase activity
	GO:0016563	Transcriptional activator activity

III. TF_{198} RELATIONAL DATA

The first step in our analysis was to compute (separate) relational data from the GO annotations data and the microarray data. The following subsections describe the methods we used to compute these relational data and we also comment on these data.

A. Gene Ontology dissimilarity data

The GO is organized as a hierarchical taxonomy of terms derived from a corpus. These terms are then used to annotate genes (or gene products) to describe the functional attributes of the genes. Pair-wise similarities between terms can be computed as in [4, 5] using shortest path and information theoretic constructs. Each gene is described by a set of terms, thus, the similarity value between two genes is some form of combination of the pair-wise term similarities. The method we use to aggregate these pair-wise term similarities is the AFMS [see 6, 7]. In brief, the AFMS is based on the Sugeno λ-measure [14]. The fuzzy densities are the importance values of each GO term in determining the similarity between two genes. Hence, important terms will have a larger impact on the overall similarity than unimportant terms. To assign a value to this importance (fuzzy density), we use the information content of each term [15]. Information content is the numerical specificity of a term—terms that are used often have a low information content, while a term that is used sparingly has a high information content [5, 15]. Terms that have a high information content are assigned respectively higher fuzzy densities. This causes them to be more important in determining the similarity between two genes, which is an intuitively pleasing result.

The AFMS overcomes limitations that are present in other similarity measures by augmenting the set of common terms between two genes with the nearest-common-ancestor of each unique term pair. This prevents a zero-valued similarity when two genes share no common terms. The following example, from [6], illustrates the AFMS calculation and provides a comparison to set-based and vector-cosine similarity measures:

Example 1 Consider two sets $G_1 = \{T_1, T_3\}$ and $G_2 = \{T_2, T_4\}$, where the associated densities (information contents) are given by the ontology shown in Fig. 2. The augmented sets are: $G_1 = \{T_1, T_3, T_6, T_5, T_7\}$ and $G_2 = \{T_2, T_4, T_6, T_5, T_7\}$, where the augmented intersection is $[G_1 \cap G_2] = \{T_5, T_6, T_7\}$. The λ-measure on G_1 is $g_1(\{T_5, T_6, T_7\}) \approx 0.26$ and on G_2 is $g_2(\{T_5, T_6, T_7\}) \approx 0.248$. Thus, the AFMS is

$$s_{AFMS}(G_1, G_2) = \frac{0.26 + 0.248}{2} = 0.25.$$

Note that a set-based similarity measure or a vector-cosine similarity measure produce a value of 0. The strength of the AFMS is that it considers the layout of the entire tree when determining gene similarity, not just the pair-wise terms.

We use the AFMS to compute a similarity for each gene pair. This results in a 198×198 similarity matrix $\mathbf{S}_{AFMS}$. We convert this to dissimilarity data by the transformation, $\mathbf{D} = [1] - \mathbf{S}$. The dissimilarity matrix is displayed in Fig. 3(a), where black is a dissimilarity value of 0.0, viz. high similarity, and white is a dissimilarity value of 1.0. Fig. 3(b) shows the *Visual Assessment of cluster Tendency* (VAT) [16] reordered dissimilarity matrix $\mathbf{D}^*_{AFMS}$. VAT reorders the dissimilarity matrix such that the cluster tendency is shown by the number of dark blocks along the diagonal. For these data, the VAT image does not show any clear cluster structure. In fact, many of the dark blocks seen in this image (such as at VAT indices 52-79 and 80-97) are genes with identical GO annotations.

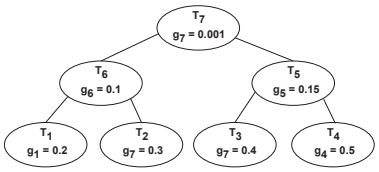
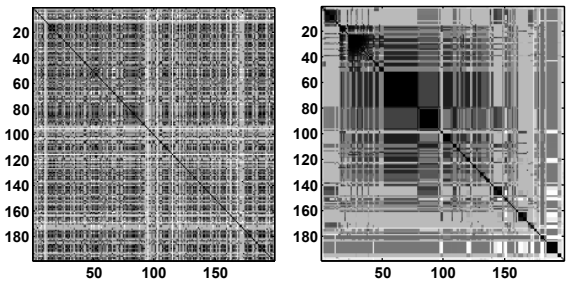


Fig. 2. Branch of an ontology with associated densities (information contents) [6]



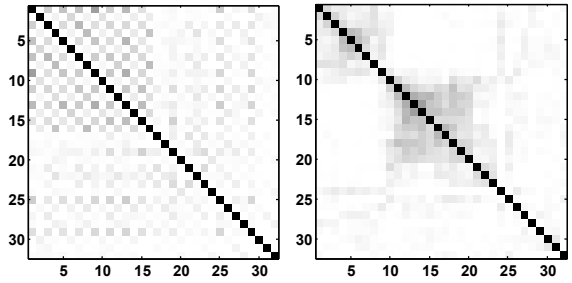
(a) Dissimilarity data - $\mathbf{D}_{AFMS}$ (b) VAT image - $I(\mathbf{D}^*_{AFMS})$

Fig. 3. AFMS-Based relational data of TF_{198} data set

B. Microarray dissimilarity data

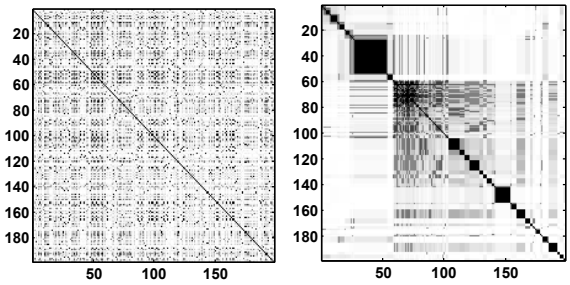
The gene expression data that we used are pre-processed tertiary data, where a value of 1 indicates *UP-regulation*, -1 indicates *DOWN-regulation*, and 0 indicates no expression. The sixteen treatments (as shown in Table I) of 3,044 differentially expressed genes are represented by a $3,044 \times 16$ tertiary matrix of expression values. We used these data to compute a 198×198 dissimilarity matrix $\mathbf{D}_{MA}$ of the TF_{198} data. The construction of this matrix is as follows.

Assume that each treatment produces no more than one of two states in a gene: UP-regulation or DOWN-regulation. 'No expression' is considered to be a NULL-hypothesis or empty state. This is due to the nature of microarray statistical analysis, which labels genes as UP or DOWN-regulated only if the expression is statistically significant. Hence, our analysis is only dependent on the statistically significant results of the microarray experiment. Since there are 16 treatments, there are 32 possible states that a gene can occupy (16 simultaneously). We computed the 32×32 similarity matrix of the states via Dice's coefficient [17]. Although many statistical similarity indices exist [see 17], we chose to use Dice's coefficient for its property of boosting the influence of matches in determining the similarity value. Consider two sets, $\{A\}$ =the set of genes in state a, and $\{B\}$ =the set of genes in state b (e.g. a is UP-regulation at six hours in local tissue fed on by *Pieris* and b is DOWN-regulation at 24 hours in systemic tissue fed on by



(a) Dissimilarity data - $\mathbf{D}_{DICE}$ (b) VAT image - $I(\mathbf{D}^*_{DICE})$

Fig. 4. Microarray insect-states DICE relational data



(a) Dissimilarity data - $\mathbf{D}_{MA}$ (b) VAT image - $I(\mathbf{D}^*_{MA})$

Fig. 5. Microarray experiment-based TF_{198} relational data

Spodoptera). For these sets, Dice's coefficient is defined as

$$s_{DICE} = \frac{2|A \cap B|}{|A| + |B|}. \tag{1}$$

Notice that the NULL-hypothesis or empty state has no effect on this similarity measure. This is an important result as the TF_{198} microarray data are very sparse. Figure 4 shows the dissimilarity data and VAT image computed with Dice's coefficient for the 32 state pairs (recall that similarity data can be converted to dissimilarity data by the transformation $\mathbf{D} = [1] - \mathbf{S}$). The checkerboard pattern in Fig. 4(a) is due to the fact that there is a low dissimilarity between UP-regulation states, even across different treatments (the same is seen between DOWN-regulation states). The figure shows that, in general, the dissimilarity values between states are very high, with the lowest off-diagonal dissimilarity value being 0.62 (a similarity of 0.38). This is a direct result of the sparsity and lack of mixing of the TF_{198} microarray data.

The VAT image, Fig. 4(b), shows two darker blocks on the diagonal, indicating that there are two clusters. Interestingly, the states that are grouped in the dark block at the top-left of the image are all UP-regulation states and the states grouped in the middle dark block are mostly caterpillar and wounding-induced DOWN-regulation states. This grouping intuitively makes sense as the chewing of the caterpillars is similar to wounding stress.

These pair-wise treatment dissimilarity values are aggregated to produce relational data for the TF_{198} data. A number of aggregation methods exist to compute the aggregated dissimilarity of two genes, including pair-wise average, normalized average, and *linear combinations of order statistics* (LOS) [see 8]. We chose the LOS operator as it has been shown to be robust to data variability and outliers [18]. The following example shows how we compute the dissimilarity value between the genes, AT5G60890 and AT2G35700, using an LOS(3) operator:

Example 2 AT5G60890 has the expression data, $\{1, 0, 0, 0, 1, 0, 0, 0, -1, 0, -1, 0, 0, 0, 0, 0\}$, where this vector is arranged in the same order as the treatment labels in Table I. AT2G35700 has the expression data, $\{0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$. There are eight unique pairs of non-zero expression data between the two genes. The LOS(3) operator will aggregate the lowest three pair-wise dissimilarity values. The dissimilarity values are $d_{1,5} = 0.62$, $d_{5,5} = 0$, $d_{9,5} = 0.99$, $d_{11,5} = 0.95$, $d_{1,6} = 0.81$, $d_{5,6} = 0.79$, $d_{9,6} = 0.93$, and $d_{11,6} = 0.93$. Thus, the dissimilarity between AT5G60890 and AT2G35700 is

$$D_{LOS(3)} = \frac{0 + 0.62 + 0.79}{3} = 0.47.$$

Figure 5 shows the dissimilarity matrix $\mathbf{D}_{MA}$ and VAT image $I(\mathbf{D}^*_{MA})$ for the TF_{198} data set computed with the LOS(3) aggregation operator. These relational data computed from the microarray experiment show that the similarity values between genes are fairly weak. Observe the dark block in the VAT image at the indices 28-53. This is a group of genes with identical expression data—each gene DOWN-regulates at six hours in the presence of Myzue feeding. We discuss this group of genes more in the next section.

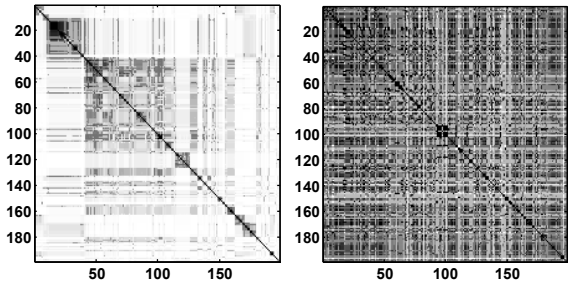
IV. Aggregate Analysis

The dissimilarity matrices, $\mathbf{D}_{GO}$ and $\mathbf{D}_{MA}$, are relational data computed from different information. In order to combine these two sources of information we aggregated the dissimilarity matrices with a fuzzy aggregation operator. Fuzzy aggregation operators are averaging operators and produce output values that are between the **MIN** and **MAX** of the input [19]. We present results for both the **MIN** and **MAX** operators. The aggregate dissimilarity matrix $\mathbf{D}_A$ is computed as

$$(D_A)_{ij} = A\left(\{(D_{GO})_{ij}, (D_{MA})_{ij}\}\right),\ i, j = 1, \ldots, 198, \tag{2}$$

where A is a chosen aggregation operator. The **MAX** operator produces a low dissimilarity only if both inputs are low, where the **MIN** operator produces a low dissimilarity if one input is low. Figure 6 shows the VAT images $I(\mathbf{D}^*)$ of the aggregated dissimilarity matrices, $\mathbf{D}_{MAX}$ and $\mathbf{D}_{MIN}$. As expected, Fig. 6 shows that the **MAX** operator produces dissimilarity values that are high and the **MIN** produces dissimilarity values that are low (recall that black is low dissimilarity $d = 0.0$, white is high dissimilarity $d = 1.0$).

We applied a relational clustering algorithm to the aggregated data. *Non-Euclidean Relational Fuzzy c-Means* (NERFcM) is a fuzzy relational clustering algorithm that is based



(a) VAT image - $I(\mathbf{D}^*_{MAX})$ (b) VAT image - $I(\mathbf{D}^*_{MIN})$

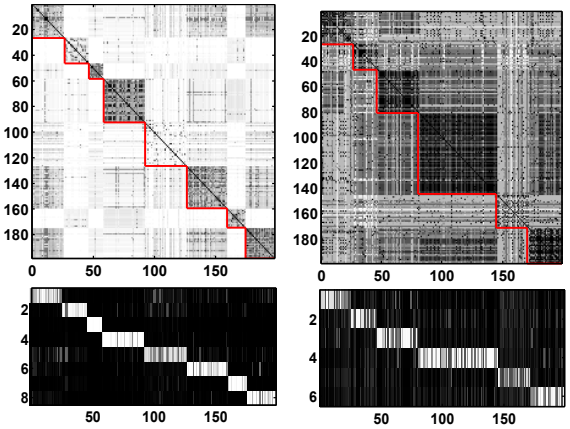
Fig. 6. Aggregated relational data for TF_{198} data set

on the fuzzy c-means clustering algorithm [11]. The parameters of NERFcM are c, the number of clusters, and m, the fuzziness index. The fuzziness index determines the fuzziness of the resultant partitions: $m = 1$ produces crisp clusters, while an $m > 1$ produces fuzzy clusters (genes can belong [partially] to more than one cluster). *Correlation Cluster Validity* (CCV) is a relational cluster validity index that has be shown to be effective with bioinformatics data [12]. We ran NERFcM for all possible pairs of $c = 2, 3, \ldots, 10$ and $m = 1.1, 1.2, \ldots, 2$. We then used CCV as a validity heuristic to determine the optimal cluster choice. Although we present results for only NERFcM and CCV, any relational clustering algorithm could be used with this data [see 20].

Figure 7 shows the "best" partition for each aggregate dissimilarity matrix. The top view shows the dissimilarity matrix ordered such that the (hardened) partitions are aligned (see [21] for a discussion of *aligned* partitions). The bottom view is the partition matrix U, where each row represents a cluster and each column is the cluster membership of a gene (white represents high membership, black is low). The optimal partitions, empirically determined using CCV, are the U_{MAX} partition produced with the parameters $c = 8$ and $m = 1.1$ and the U_{MIN} partition produced with the parameters $c = 6$ and $m = 1.4$. Notice that the U_{MIN} partition is fuzzier than the U_{MAX} partition, which is intuitively accurate as there is overall lower dissimilarity between the genes in the U_{MIN} partition (see Figs. 6 and 7).

A. Cluster summary analysis

Functional summarization is very important as it gives bioinformatics researchers information to determine the function of lesser-studied genes. We perform summarization by computing the frequency of each GO annotation within a cluster of genes [22]. If most or all the genes within a cluster are annotated by the same term, we infer that this term is a good summarization. However, there is a downside to this simple counting method. The summarizing term is often very general. This is seen in the TF_{198} data set, where almost every gene is annotated by the GO term: *transcription factor activity*. To circumvent this problem, we use an information-content weighted functional summarization method [22]. The summarization is determined by weighting each term by its information content and then computing the *most-representative-terms* (MRT) with a weighted ranking. Terms that are used less often have a higher information content. The weighting method produces more *specific* and hence, we argue, more useful summarizations.



(a) Partition - U_{MAX}) (b) Partition - U_{MIN})

Fig. 7. NERFcM (with CCV) partition data for TF_{198} data set

TABLE III
FUNCTIONAL SUMMARIZATION OF SELECTED TF_{198} CLUSTERS IN RELATIONAL DATA $\mathbf{D}_{MAX}$

Cluster No.	**3**	**6**
N_G	12	33
MRTs (GO)	TF activity (9) Nucleus (6) R. to salt stress (2)	TF activity (29) Nucleus (21) Reg. of Trans., DNA-dep.
MRTs (MA)	*Myzus* UP 6hr-L (12) *Pieris* UP 24hr-S (1)	*Brevicoryne* UP 24hr-L (23) *Spodoptera* UP 24hr-S (22) *Spodoptera* UP 24hr-L (10)

Tables III and IV display the MRTs for selected clusters in the partitions shown in Fig 7. We also show the most frequent microarray states. Note that each of the MRTs has a parenthetical notation that shows its count in the cluster. These values are important as researchers can examine the summarizations of each cluster and determine relations between microarray treatments and gene functions.

Consider clusters **3** and **6** in U_{MAX} (Fig. 7(a)). Table III shows that all twelve genes in cluster **3** are UP-regulated at six hours in local tissue in the presence of *Myzus* feeding. Furthermore, the GO-annotations show that two of these genes are *responsive to salt stress*. This table also shows that 23 of the 33 genes in cluster **6** are UP-regulated in the presence of *Brevicoryne* feeding and 22 are UP-regulated in the presence of *Spodoptera* feeding (in both local and systemic tissue). Interestingly, *Brevicoryne* is an aphid and *Spodoptera* is a caterpillar, two distinctly different feeding mechanisms.

Now consider clusters **1** and **6** in U_{MIN} (Fig. 7(b)). Table IV shows that 18 of the 26 genes are annotated by the GO term: *zinc ion binding*, and that nearly a third of the 26 genes are DOWN-regulated in the presence of *Myzus* feeding. Finally, we examine cluster **6** in U_{MIN}. The GO MRTs

TABLE IV
FUNCTIONAL SUMMARIZATION OF SELECTED TF_{198} CLUSTERS IN RELATIONAL DATA $\mathbf{D}_{MIN}$

Cluster No.	**1**	**6**
N_G	26	28
MRTs (GO)	Zinc ion binding (18)	R. to salt stress (17)
	TF activity (23)	R. to Jasmonic acid (13)
	Reg. of trans. (16)	R. to Auxin (15)
MRTs (MA)	*Myzus* DN 6hr-L (7)	*Myzus* DN 6hr-L (7)
	Spodoptera UP 24hr-L (4)	*Spodoptera* UP 24hr-S (5)
	Brevicoryne UP 24hr-L (3)	Wounding UP 6hr-L (5)

of this cluster are all related to stress responses and seven of these genes DOWN-regulated in the presence of *Myzus* feeding. Interestingly, the *Myzus*-related treatments caused a larger number of genes to express than any other treatment. While we hesitate to draw concrete conclusions about these specific data, we believe that the examples shown here provide evidence that our tools are effective for investigating relations between microarray treatments and gene functions.

V. FUTURE WORK

In the future, we will further exemplify the strengths of using multiple sources of information in bioinformatics analysis by leveraging the semantic relationships in the microarray experiment to produce additional relational data. For example, we have prior knowledge of the four insects and the method in which they feed—the *Pieris* and *Spodoptera* are both caterpillars and the *Brevicoryne* and *Myzus* are aphids. Additionally, it is known that *Pieres* and *Brevicoryne* are specialists—they usually feed on Arabidopsis—and *Spodoptera* and *Myzus* are generalists, meaning they will feed on just about anything. We think that including these facts about the insects will reveal further information about the genes' functions.

Bioinformatics data tend to be very sparse; thus, tying together multiple sources of information is important to creating a clear picture of the relationships that exist within and between genes and gene products.

REFERENCES

[1] S. Khan, G. Situ, K. Decker, and C. Schmidt, "GoFigure: Automated Gene Ontology annotation," *Bioinf.*, vol. 19, no. 18, pp. 2484–2485, 2003.

[2] E. Kretchmann, W. Fleischmann, and R. Apweiler, "Automatic rule generation for protein annotation with the c4.5 data mining algorithm applied to SWISS-PROT," *Bioinf.*, vol. 17, no. 10, pp. 920–926, 2001.

[3] The Gene Ontology Consortium, "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Res.*, vol. 32, pp. D258–D261, 2004.

[4] P. Lord, R. Stevens, A. Brass, and C. Goble, "Semantic similarity measure as a tool for exploring the gene ontology." *Pacific Symposium on Biocomputing*, pp. 601–612, 2003.

[5] J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical ontology." *Proc. of Int. Conf. Res. on Comp. Linguistics X*, 1997.

[6] J. Keller, M. Popescu, and J. Mitchell, "Taxonomy-based soft similarity measures in bioinformatics," in *Proc. IEEE Int. Conf. on Fuzzy Systems*. Budapest, Hungary: IEEE, July 2004, pp. 23–30.

[7] M. Popescu, J. Keller, and J. Mitchell, "Fuzzy measures on the Gene Ontology for gene product similarity," *IEEE Trans. on Computational Biology and Bioinformatics*, vol. 3, no. 3, pp. 263–274, 2006.

[8] J. Keller, J. Bezdek, M. Popescu, N. Pal, J. Mitchell, and J. Huband, "Gene ontology similarity measures based on linear order statistics," *Int. J. on Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 14, no. 6, pp. 639–661, 2006.

[9] P. Baldi and G. Hatfield, *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*. Cambridge University Press, 2002.

[10] E. Blalock, *A Beginner's Guide to Microarrays*. New York: Kluwer Academic Publishers, 2003.

[11] R. Hathaway and J. Bezdek, "NERF c-MEANS: Non-euclidean relational fuzzy clustering," *Pattern Recognition*, vol. 27, pp. 429–437, 1994.

[12] M. Popescu, J. Bezdek, J. Keller, T. Havens, and J. Huband, "A new cluster validity measure for bioinformatics relational datasets," *to appear in Proc. IEEE WCCI*, June 2008.

[13] D. Swarbreck et al., "The Arabidopsis Information Resource (TAIR): gene structure and function annotation," *Nucleic Acids Res.*, vol. 36, pp. D1009–D1014, 2008.

[14] M. Sugeno, *Fuzzy Automata and Decision Processes*. New York: North-Holland, 1977, ch. Fuzzy measures and fuzzy integrals: a survey, pp. 89–102.

[15] P. Resnik, "Semantic similarity in a taxonomy: an information-base measure and its application to problems of ambiguity in natural language," *J. Artifical Intelligence Research (JAIR)*, vol. 11, pp. 95–130, 1999.

[16] J. Bezdek and R. Hathaway, "VAT: A tool for visual assessment of (cluster) tendency," in *Proc. IJCNN 2002*, Piscataway, NJ, 2002, pp. 2225–30.

[17] C. van Rijsbergen, *Information Retrieval*. London: Butterworths, 1979.

[18] J. Hosking, "L-moments: analysis and estimation of distributions using linear combinations of order statistics," *J. of the Royal Statistical Society, Series B*, vol. 52, pp. 105–124, 1990.

[19] G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Upper Saddle River, New Jersey: Prentice Hall, 1995.

[20] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[21] T. Havens, J. Bezdek, J. Keller, and M. Popescu, "Clustering in ordered dissimilarity data," *in review, Pattern Recognition*, 2008.

[22] M. Popescu, J. Keller, J. Mitchell, and J. Bezdek, "Functional summarization of gene product clusters using Gene Ontology similarity measures," *Proc. 2004 ISSNIP*, pp. 553–559, 2004.